2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Superpixel Appearance and Motion Descriptors for Action Recognition

Xuan Dong
Faculty of Information Technology
Macau University of Science and Technology, Avenida Wai Long
Taipa, Macau
Email: dongxuanowen@gmail.com

Ah-Chung Tsoi
Faculty of Information Technology
Macau University of Science and Technology, Avenida Wai Long
Taipa, Macau
Email: actsoi@must.edu.mo

Sio-Long Lo
Faculty of Information Technology
Macau University of Science and Technology, Avenida Wai Long
Taipa, Macau
Email: sllo@must.edu.mo

***Abstract*—This paper introduces a novel video representation based on superpixel segmentation and appearance and motion descriptors. Superpixel represents a very useful preprocessing step for a wide range of computer vision applications, as they group pixels into perceptually meaningful atomic regions which can be used for recognizing complex motion patterns. We construct a novel video representation in terms of superpixel-based histograms of oriented gradients (HOG), histograms of optical flow (HOF) and motion boundary histograms (MBH) descriptors, and integrate such representations with a bag-of-features (BoF) model for classification. The proposed approach is evaluated in the context of action classification on a challenging benchmark dataset: UCF Sports dataset and it achieves 87.9% generalization accuracy. The experimental results demonstrate the advantage of superpixel-based descriptors compared to other approaches for human action recognition.**

I. Introduction

Human action recognition in videos has been a very active research area in computer vision and machine learning over the past years, as it can be considered as one of the key pre-requisites for video analysis and understanding, such as video indexing and retrieval, human-computer interaction, and activity monitoring [1]–[3]. In general, a popular approach is to extract a set of local descriptors first, and then to apply a bag-of-features model [5] for matching these local descriptors obtained in the set of training video clips, which are labeled, to those yet unlabeled in the testing set. Despite recent developments, the representation of local regions in videos is still an open field of research.

Although much research has been reported on human action classification, recognizing human actions from realistic videos still presents a challenging problem due to the significant camera motion, background clutter and changes in object appearance, scale and illumination conditions [1], [4], [6], [7]. There are many existing methods for representing a video, for example, histograms of oriented gradients (HOG) [8], histograms of optical flow (HOF) [9], motion boundary histograms (MBH) [10], trace transform [4] etc. This is because each primitive (descriptor) cannot fully capture the underlying system dynamics of human action as provided through pixel intensity variations in space and time in the video. The idea to utilize superpixels as a primitive for image analysis and processing was introduced in [11]. Superpixel is a region in which neighboring pixels with similar low-level features like color or texture are grouped into perceptually meaningful homogeneous region, which creates a spatial support for region-based features [12]. Regions [13] obtained by superpixels might be a more natural representation of reality and an object is usually composed of several superpixels. Meanwhile, superpixel provides a compact representation of an original image, which has a great improvement in computational efficiency, lower run time and memory cost [14]. Especially for video applications [1], [6], [12], the usage of superpixels instead of raw pixel data is proposed, as otherwise a vast amount of data has to be handled. In this paper, the basic idea is to consider each superpixel as a single entity. Then a new approach to construct superpixel appearance and motion descriptors is proposed, and it can be used as one way of video representation. Once a good set of descriptors is obtained, the learning would be straightforward, applying standard classification algorithms like support vector machines, or kernel machines.

The remainder of the paper is organized as follows: Section II presents related work. In Section III, we briefly summarize the generation of SLIC superpixel algorithm, and then explain our approach for superpixel-based HOG, HOF and MBH descriptors respectively. The classification framework is presented in Section IV, and experimental results and comparisons with other state-of-the-art approaches are shown in Section V. Some conclusions are given in Section VI.

II. Related work

Local space-time features [4], [15] provide a good video representation for human action recognition. Such features are usually extracted directly from videos, and they capture appearance and motion characteristics and provide relatively independent representation of sequences with respect to their spatio-temporal shifts and scales as well as background clutter and multiple motion in the scene [3]. There are a number of approaches for extracting local space-time features in videos [4], [15]–[17]. The Harris3D detector [15] introduces spatio-temporal interest points which are local maxima of a cornerness condition at each pixel. The Cuboid detector [16] is based on a cornerness function which combines a 2D Gaussian filter in space with a 1D Gabor filter in time. The Hessian detector [17] measures the saliency with the determinant of the Hessian matrix. Trace transform, a generalization of Randon transform can be used to extract features]citetrace. Dense

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

sampling [3], [18] extracts video patches at regular positions and scales in space and time. To describe space-time features, feature descriptors are introduced to capture shape and motion in the neighborhoods of selected points using image measurements, which include higher order derivatives [19], brightness information [16], HOG [8], HOF [9] and MBH [10], to spatio-temporal extension of image descriptors, such as 3D-SIFT [20], HOG3D [2] and extended SURF [17].

Tracking interest points [7] through video sequences is an alternative approach to handle different characteristics of space domain and time domain in videos. Spatio-temporal interest points encode information at a given location in space and time. In contrast, trajectories track a given spatial point over time in video sequences and capture motion information. To obtain feature trajectories, either a KLT (Kanade-Lucas Tracker) is used [21], [22] or SIFT descriptors between consecutive frames are matched [23], [24]. Recently, a few publications [7], [25], [26] show good performances in a number of benchmark datasets using a dense trajectories approach.

Since the introduction of the concept of superpixel, it becomes much more efficient to process high level representation in many vision applications [27]–[29] than using pixels. In general, algorithms for generating superpixels can be categorized into graph cuts approaches [27], [30]–[32], generative models [29] and clustering approaches [33]–[36]. Graph cuts approaches treat each pixel as a node and the similarity between neighboring pixels using an edge weight. Superpixels are obtained by minimizing a global cost function. Generative models [29], on the other hand, assumes that superpixels are generated by an underlying Bayesian model. Clustering approaches group pixels together into regions by performing a gradient ascent process iteratively. The clusters formed are gradually refined until convergence to obtain the superpixels. In this paper, the *Simple Linear Iterative Clustering* (SLIC) [14] is applied to generate superpixels due to its simplicity and efficiencies in contrast to other state-of-the-art methods [28], [29], [34].

III. Superpixel-based descriptors

In this section, we present a way on how to extract superpixel-based descriptors from video clips. An overview of the SLIC algorithm is shown in Section III-A. Then a presentation of the appearance and motion descriptors is provided in Section III-B and Section III-C respectively.

A. SLIC superpixel segmentation

SLIC segmentation [14] is an efficient cluster technique, which bears some resemblance to the *k*-means clustering method. Pixels of an image are considered as data points in a multi-dimensional feature space in which each dimension corresponds to a color channel or image coordinate of the pixels under consideration. Superpixels are represented by clusters in this multi-dimensions feature space and each data point can only be assigned to one cluster. This assignment determines the segmentation and the superpixel is obtained. SLIC segmentation technique can produce superpixels with uniform size and shape, maintain connectivity and compactness, and preserve original image edges [14], in contrast to those formed by using a generative model [29], which may allow the shape and size of superpxiels to vary. As shown in Fig. 1, the lattice-like superpixels are obtained, which are highly homogenous and compact.

(a) Diving. (b) Lifting.

(c) Swinging on the pommel horse. (d) Swinging at the high bar.

Fig. 1. Illustration of segmented UCF Sports images by SLIC algorithm. The red dot points are cluster centers when the superpixel size equals to 30×30 pixels approximately.

Specifically [14], given an image with N pixels, it starts with sampling K regularly spaced cluster centers at every grid step $S = \sqrt{N/K}$, and perturbing them in a $Q \times Q$ neighborhood to the lowest gradient position, $Q = 3$ in our experiments. Then clustering is performed in a five dimensional [*xylab*] space, where [*lab*] is the pixel color vector in CIELAB color space and *xy* is the pixel position. Pixels in the input image p_i are assigned to the best matching cluster center c_k from a $2S \times 2S$ neighborhood around the cluster center by minimizing the distance measure D_{total}

$$D_{\text{total}}(p_i, c_k) = D_{\text{color}}(p_i, c_k) + \frac{m}{S} D_{\text{space}}(p_i, c_k),$$

where D_{color} is the color distance which ensures the superpixel is homogeneous, D_{space} is the spatial distance which forces the compactness of superpixel, and m controls the relative weight between color similarity and spatial proximity. The greater the value of m, the more spatial proximity is emphasized and the more compact the cluster. $m = 10$ is chosen in this paper. This roughly matches the empirical maximum perceptually meaningful CIELAB distance and offers a good balance between color similarity and spatial proximity. After several iterations of the above operation, the algorithm converges to the best position of the cluster centers. Further details can be found in the supplement of [14].

B. Appearance descriptor

The basic idea of an appearance descriptor is that local object appearance can often be characterized rather well by the distribution of local intensity gradients. Histograms of oriented gradients (HOG) descriptor [8] is one of the appearance descriptors for representing visual appearance of images, which provides a robust feature set that allows the human form to

be discriminated clearly, even in cluttered backgrounds and changed illuminations.

In order to compute a histogram of superpixel-based gradient orientations, the superpixel can be considered as a pixel, i.e., the cluster center of SLIC can be considered as a superpixel representation. The rationale for this is that once the clustering converged, the cluster center together with its neighborhood centers contain most information necessary to characterize an action. A superpixel acts in the same manner as the cell, and its adjacent superpixels and itself behave like the block in pixel-based HOG. At each superpixel, the image gradient vector is calculated and converted to an angle, voting into the corresponding orientation bin with a vote weighted by the gradient magnitude. Votes are accumulated over this superpixel block and a L_2 normalization runs on each superpixel block to provide strong illumination invariance. The steps involved in obtaining the superpixel based HOG are illustrated in Fig. 2. In order to simplify the description, a circle is used to represent superpixel in Fig2.

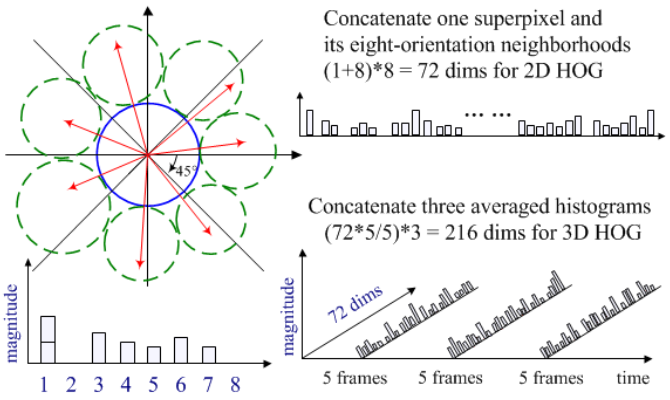


Fig. 2. Illustration of the computation of superpixel-based HOG descriptor. (a) Gradient vectors among the middle superpixel and its adjacent superpixels need to be computed (upper left); (b) An 8-bin histogram of gradient orientations is computed over a set of gradient vectors (bottom left); (c) A 72-dimensional histogram for 2D HOG is a result of concatenation of one superpixel and its eight-orientation neighborhoods. Eight-orientation neighborhoods means that the *xy* plane is divided into eight equal parts, and the cluster center of superpixel falling into a certain part indicates that the superpixel belongs to the corresponding orientation. If there is no superpixel in some orientations, an 8-bin zero histogram will be padded (upper right); (d) An averaged 2D HOG is computed between five frames and the 3D HOG is a concatenation of three averaged 2D HOG (bottom right).

Empirically, the length of the construction along the video sequences is set to 15 frames (i.e., 3×5 frames).

C. Motion descriptors

For encoding the local motion information, histograms of optical flow (HOF) [9] are employed to analyze the motion of the objects in an image. In practice, the implementation of superpixel-based HOF is similar to that of HOG, but the gradient $g = (\mathcal{I}_x, \mathcal{I}_y)$ in HOG is replaced by the optical flow field $\omega = (\mathcal{I}^u, \mathcal{I}^v)$ used for HOF, where $\mathcal{I}_x$ and $\mathcal{I}_y$ are the x and y derivative of image $\mathcal{I}$ respectively, and $\mathcal{I}^u$ and $\mathcal{I}^v$ are the horizontal and vertical optical flow components. For the 2D superpixel-based HOF descriptor, an additional zero bin is added which accounts for the optical flow magnitudes when they are lower than a threshold. The dimension of the final HOF descriptor is 243 (i.e., 81×3).

Optical flow represents the absolute motion between two frames. We need another motion descriptor which characterizes human action well while remaining resistant to typical camera and background motions. Motion boundary histograms (MBH) descriptor [10] computes the derivatives separately for the horizontal and vertical components of the optical flow, which encodes the relative motion between pixels and removes the locally constant camera motion. For this reason, MBH is much more robust and discriminative for action recognition [7].

The simplest approach to obtain a superpixel-based MBH descriptor is to treat horizontal and vertical components of the optical flow $\omega = (\mathcal{I}^u, \mathcal{I}^v)$ as independent. Spatial derivatives $\{(\mathcal{I}^u_x, \mathcal{I}^u_y), (\mathcal{I}^v_x, \mathcal{I}^v_y)\}$ are computed for each of them, and the orientation information is quantized into 9-bin histograms and the magnitude is used for vote weighting for each component (i.e., MBHx and MBHy) in the same manner as those of the HOF descriptor. Both histogram vectors are normalized separately with their L_2 norms. For both MBHx and MBHy the feature vector size is 243 (i.e., 81×3).

IV. Classification

In this section, we will first describe the bag-of-features approach as used in our experiments and then we will describe the dataset and the experimental protocol.

A. Bag-of-features model

The standard bag-of-features (BoF) [19], [37]–[39] approach is applied to convert the local descriptors from a video into a fixed dimensional vector. First, a codebook for each descriptor is constructed separately by the *k*-mean clustering algorithm, and then the clusters will be served as visual words [40]. Descriptors are then assigned to their closest visual word using the Euclidean distance. This is essentially a vector quantization step in an unsupervised learning approach. The resulting histograms of visual word occurrences are used as video representations.

For classification we use a kernel machine with a χ^2 kernel:

$$K_{\chi^2}(H_i, H_j) = exp(-\frac{1}{2A}\sum_{k=1}^{V}\frac{(h_{ik} - h_{jk})^2}{h_{ik} + h_{jk}})$$

where $H_i = \{h_{ik}\}_{k=1}^{V}$ and $H_i = \{h_{jk}\}_{k=1}^{V}$ are the frequency histograms of visual word occurrences and V is the vocabulary size. A is the average channel distance between all training samples [41]. In the case of multi-class classification, the one-against-one SVM is applied. It constructs $M \times (M-1)/2$ binary classifiers, using all the binary pair-wise combinations of the M classes. Each classifier is qualified by using the examples of the first class as positive and the examples of the second class as negative examples. To combine these classifiers, the Max Wins algorithm is used. It finds the subsequent class by selecting the class voted by the majority of the classifiers.

Typically, the approach for integrating the contribution of different descriptors is the multi-channel SVM, which is a case of multiple kernel learning. We simply average the

kernels computed from different representations to combine different channels using the idea of multi-channel SVM. In the following, we use the symbol "+" to represent combined descriptors, for example, "HOG+HOF" and "MBHx+MBHy".

B. Dataset

The UCF Sports dataset [42] which contains ten actions: diving, golf swinging, kicking a ball, weight lifting, horse riding, running, skateboarding, swinging (on the pommel horse and on the floor) and swinging (at the high bar). The dataset consists of 150 video clips which are taken from real sports broadcasts. We follow the same train/test samples protocol proposed in [43], in which one third of videos from each action category is taken to form the test set, and the rest of the video clips are used for training purposes.

V. Experimental results

This section evaluates superpixel appearance and motion descriptors (HOG, HOF and MBH) on the UCF Sports dataset. The kernel machine with a χ^2 kernel is employed for classification. The experiments are run 10 times for various parameters and the average accuracy precision (AAP) [39] are reported. In this paper, $SP = n$ denotes that the size of superpixel is $n \times n$ pixels approximately. The parameters of our approach are discussed in Section V-A. Section V-B presents the performance of different superpixel sizes for various descriptors. Finally, we compare our results with those obtained using state-of-the-art approaches in Section V-C.

A. Parameter learning

For the construction of codebook in BoF model, a few publications [2], [3], [7] argue that fixing the number of visual words per descriptor to 4000 will give good results for a wide range of datasets. However, in the superpixel-based case as shown in Table I, HOG, HOF, MBHx and MBHy all give the best results in the range of 400 and 500. The reason is that superpixels are the results of perceptual grouping of pixels and they carry more information than pixels, so the number of visual words is reduced. Meanwhile, the usage of superpixel instead of raw pixel data has a great improvement in computational efficiency, as otherwise a vast amount of data has to be handled. In the following experiments, we report the best AAP between 400 to 600 visual words and cluster a subset of 15000 randomly selected training features using *k*-means clustering technique.

TABLE I. Results of various sizes of visual words ranging from 400 to 4000 on HOG, HOF, MBHx and MBHy descriptors ($SP = 25$). The best results for each descriptor are shown in bold.

	HOG	HOF	MBHx	MBHy
400	71.9	**79.6**	84.0	82.1
500	**74.6**	77.1	**84.2**	**82.9**
600	73.0	76.5	83.8	82.3
800	71.6	76.1	81.9	80.0
1000	69.9	75.3	77.4	77.3
2000	65.9	72.9	76.5	76.1
3000	64.7	69.9	75.1	74.5
4000	63.8	65.8	72.5	70.9

B. Evaluation of superpixel size

We report seven different SPs on the UCF Sports dataset in Table II. The best results for each descriptor are almost always given by $SP = 20$ except in one case (MBHy). Especially, superpixel-based HOF gives surprisingly good results by itself and it achieves the highest accuracy 87.9% in our experiments. The second best result 86.7% is obtained by the combination of HOG and HOF. Separate HOF outperforms HOG due to the intuitive fact that motion is more discriminative than static appearance for action recognition. Generally, MBHx and MBHy show similar performance, and both of them obtain a relatively higher accuracy than other descriptors on various superpixel size except $SP = 20$. Fig. 3 gives the confusion tables of different superpixel descriptors.

TABLE II. Comparison of different superpixel-based descriptors on the UCF Sports dataset. Images are segmented into $SP = 32, 30, 25, 22, 20, 18, 15$ by SLIC [14]. The symbol '+' denotes that integrating the contribution of different descriptors with the multi-channel SVM. "Combined" indicates that all descriptors (HOG, HOF, MBHx and MBHy) are combined using the multi-channel approach. The best results for each descriptor are shown in bold.

	32	30	25	22	20	18	15
HOG	71.5	72.9	74.6	78.1	**80.0**	77.8	77.3
HOF	75.7	79.1	79.6	82.9	**87.9**	82.8	80.5
HOG+HOF	77.2	81.1	81.3	84.4	**86.7**	84.7	80.3
MBHx	80.6	83.4	84.2	84.8	**85.1**	**85.1**	82.3
MBHy	79.4	81.7	82.9	84.9	84.9	**85.1**	81.9
MBHx+MBHy	81.2	81.9	82.3	**85.1**	**85.1**	**85.1**	83.0
Combined	80.3	82.1	83.4	83.9	**84.3**	83.9	81.2

Table II shows that the choice of superpixel size is a key issue for performance. On one hand, using large superpixel may bring in the risk of a superpixel spanning across multiple semantic objects. On the other hand, a small superpixel may contain insufficient points to precisely define a good features.

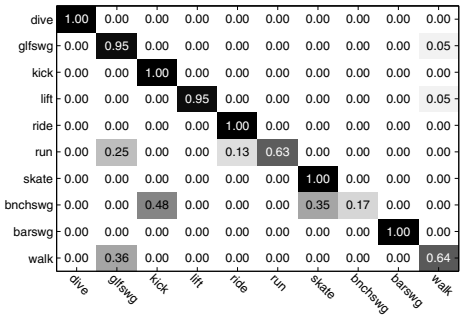
C. Comparison with state-of-the-art approaches

First, we compare superpixel-based descriptors (HOG, HOF, HOG+HOF and MBHx+MBHy) with a few pixel-based descriptors separately in Table III, and comparison of the superpixel-based descriptors to state-of-the-art approaches is shown as Table IV.

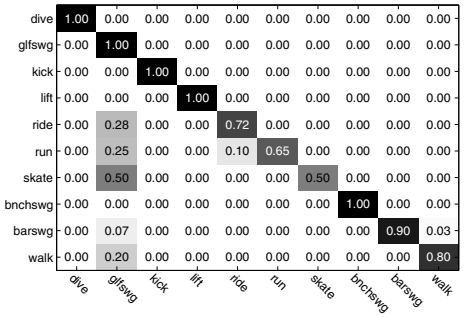
TABLE III. Comparison of superpixel-based descriptors with pixel-based descriptors. The first five rows are the results of local space-time features approaches and the next three rows are those of trajectories approaches. Average accuracy over all classes are reported on the UCF Sports dataset. MBH is the combination of MBHx and MBHy.

	HOG	HOF	HOG+HOF	MBH
Harris3D [15]	71.4	75.4	78.1	–
Cuboids [16]	72.7	76.7	77.7	–
Hessian [17]	66.0	75.3	79.3	–
Dense sampling [3]	77.4	82.6	81.6	–
Dense cuboids [7]	80.2	77.8	–	83.2
SIFT trajectory [23]	74.2	69.9	–	72.1
KLT trajectory [21]	80.2	72.7	–	78.4
Dense trajectory [7]	**84.3**	76.8	–	84.2
Superpixel-based	80.0	**87.9**	**86.7**	**85.1**

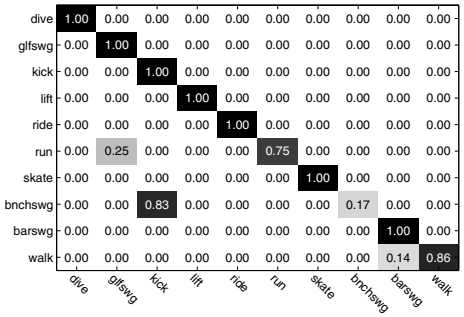
For HOG case, the best results is obtained by dense trajectory [7]. Because UCF Sports dataset is sports related and



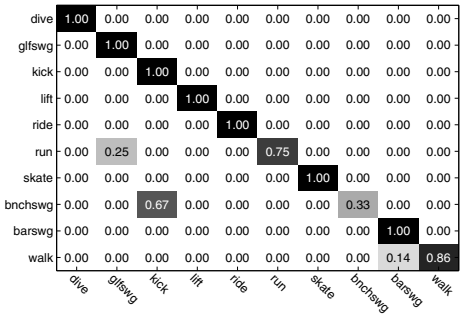
(a) Confusion table of superpixel-based HOG.



(b) Confusion table of superpixel-based HOF.



(c) Confusion table of superpixel-based MBH.



(d) Combined descriptors (HOG+HOF+MBH).

Fig. 3. Confusion tables of superpixel-based descriptors on the UCF Sports dataset, and χ^2 kernel SVM is applied.

the spatial context is very informative for sports actions as they often involve specific equipment and scene types. Meanwhile, HOG is designed to encode the static context information, and dense trajectories capture the background which may provide useful context information. However, for HOF, HOG+HOF and MBH cases, superpixel motion descriptors outperform the other approaches significantly. Superpixel-based HOF and MBH descriptors improve the results as they represent zero-order (HOF captures the absolute motion between two frames) and first order (MBH encodes relative motion between pixels in the optical flow) motion information.

TABLE IV. COMPARISON OF THE SUPERPIXEL-BASED DESCRIPTORS TO STATE-OF-THE-ART APPROACHES, AS REPORTED IN THE CITED PUBLICATIONS.

UCF Sports	
Wang *et. al.* [3]	85.6
Kläser *et. al.* [44]	86.7
Kovashka *et. al.* [46]	87.3
Le *et. al.* [45]	86.5
Wang *et. al.* [7]	88.0
Superpixel-based	**87.9**

Table IV compares the proposed approach to state-of-the-art results on the UCF Sports dataset. We can observe that superpixel-based descriptors improve over these four pixel-based local space-time features approaches [3], [44]–[46] and are comparable to dense trajectories approach [7]. The encouraging results illustrate the ability of our method to extract effective features for classification using superpixel appearance and motion descriptors. The improvement of our descriptors is also substantial as the number of features used for training and testing is decreased and it is faster and more memory efficient under the framework of BoF model.

VI. CONCLUSIONS

In this paper, a novel approach for efficient video representation based on superpixel segmentation technique are proposed. Superpixel segmentation preserves the salient features of a pixel-based representation. Superpixel appearance and motion descriptors are designed to be robust to the significant intra-class variations, occlusion and background cluster. In the experimental results, it is found that our approach has been shown to outperform previous approaches which extract local features on the common pixel-based representation. It is also indicated that superpixel size is a critical factor in the descriptor performance. The experiments verified that superpixel-based video representation is effective for recognizing the natural and realistic actions, and using hybrid features of appearance and motion can improve the average recognition accuracy.

In future work, we wish to investigate additional cues or features for describing superpixels, i.e., edge information and spatial location information. Meanwhile, we will consider the temporal link between superpixels in successive images, e.g., the same image regions in consecutive frames are consistent. We believe that temporal superpixels will bridge the relationship between superpixels and videos.

Acknowledgements. This work was financially supported by Fundo para o Desenvolvimento das Ciências e da Tec-

nologia, Macau Special Administrative Region, China. Grant Number: 034/2011/A2.

REFERENCES

[1] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild"', *CVPR*, 2009.

[2] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," *BMVC*, 2008.

[3] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *BMVC*, 2009.

[4] G. Goudelis, K. Karpouzis, and S. Kolhas. "Exploring trace transform for robust juman action recognition;;. *Pattern Recognition*, vol. 46, no. 12, pp. 3238–3248, 2013.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," *ECCV Workshop*, vol. 1, pp. 22, 2004.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," *ICCV*, pp. 2556–2563, 2011.

[7] H. Wang, A. Kläser, C. Schmid and, C-L Liu, "Action recognition by dense trajectories", *CVPR*, 2011.

[8] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, 2005.

[9] I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," *CVPR*, 2008.

[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *ECCV*, 2006.

[11] X. Ren, J. Malik, "Learning a classification model for segmentation," *ICCV*, 2003.

[12] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image", *ICCV*, 2005.

[13] T. Malisiewicz, A. Efros, "Improving spatial support for objects via multiple segmentations", *BMVC*, 2007.

[14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[15] I. Laptev, "On space-time interest points", *IJCV*, 2005.

[16] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features", *ICCCN*, 2005.

[17] G. Willems, T. Tuytelaars, and L. J. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector, " *ECCV*, 2008.

[18] E. Nowak, F. Jurie and B. Triggs, "Sampling strategies for bag-of-features image classification" *ECCV*, 2006.

[19] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," *ICRP*, 2004.

[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SITF descriptor and its application to action recognition," *ACM Conference on Multimedia*, 2007.

[21] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features," *ICCV Workshop*, 2009.

[22] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," *CVPR*, 2009.

[23] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," *CVPR*, 2009.

[24] J. Sun, Y. Mu, S. Yan, and L. Cheong, "Activity recognition using dense long-duration trajectories," *ICME*, 2010.

[25] J. Wu, D. Hu, "Learning effective event models to recognize a large number of human actions" *IEEE Transactions on Multimedia*, 2013.

[26] K. Avgerinakis, A. Briassouli and I. Kompatsiaris, "Recognition of Activities of Daily Living" *ICTAI*, 2012.

[27] P. F. Felzenszwalb, D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, 2004.

[28] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa, "Efficient hierarchical graph-based video segmentation," *CVPR*, 2010.

[29] T. H. Thi, J. Zhang, L. Cheng, L. Wang, and S. Satoh, "Human action recognition and localization in video using structured learning of local space-time features," *AVSS*, pp. 204–211, 2010.

[30] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," *ECCV*, 2010.

[31] A. Moore, S. Prince, J. Warrell, U. Mohammed, and G. Jones, "Superpixel lattices," *CVPR*, 2008.

[32] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[33] M. Van Bergh, X. Boix, G. Roig, B. Capitani and L. Van Gool, "SEEDS: superpixels extracted via energy-driven sampling," *ECCV*, pp. 13–26, 2012.

[34] A. Levinshtein, A. Stern, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi, "TurboPixels: fast superpixels using geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[35] A. Vedaldi, S. Soatto, "Quick shift and kernel methods for mode seeking," *ECCV*, 2008.

[36] D. Comaniciu, P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[37] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," *ICCV*, pp. 1–8, 2007.

[38] J. C. Niebles, H. Wang, and L. Fei-fei, "Unsupervised learning of human action categories using spatial-temporal words," *BMVC*, 2006.

[39] M. Marszałek and I. Laptev and C. Schmid, "Actions in context," *CVPR*, pp. 2929–2936, 2009.

[40] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos", *ICCV*, pp. 1470–1477, 2003.

[41] J. Zhang, M. Marszałek, S. Lazebnik and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.

[42] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach: A spatio-temporal maximum average correlation height filter for action recognition," *CVPR*, 2008.

[43] T. Lan, Y. Wang, W. Yang, N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[44] A. Kläser, M. Marszalek, I. Laptev, and C. Schmid, "Will person detection help bag-of-features action recognition?" *INRIA: Research Report*, 2010.

[45] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *CVPR*, 2011.

[46] A. Kovashka, K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," *CVPR*, 2010.